Journal
Association of Applied Biologists
Society for Experimental Biology

Plant Biotechnology Journal (2012) **10**, pp. 703–708 doi: 10.1111/j.1467-7652.2012.00717.x

Bread matters: a national initiative to profile the genetic diversity of Australian wheat

David Edwards[1], Stephen Wilcox[2], Roberto A. Barrero[3], Delphine Fleury[4], Colin R. Cavanagh[5,6], Kerrie L. Forrest[7], Matthew J. Hayden[7], Paula Moolhuijzen[3], Gabriel Keeble-Gagnère[3], Matthew I. Bellgard[3], Michał T. Lorenc[1], Catherine A. Shang[8], Ute Baumann[4], Jennifer M. Taylor[5], Matthew K. Morell[5], Peter Langridge[4], Rudi Appels[3] and Anna Fitzgerald[8,*]

[1]*Australian Centre for Plant Functional Genomics and University of Queensland, St. Lucia, Qld, Australia*
[2]*Australian Genome Research Facility, The Walter and Eliza Hall Institute of Medical Research, Parkville, Vic., Australia*
[3]*Centre for Comparative Genomics, Murdoch University, Perth, WA, Australia*
[4]*Australian Centre for Plant Functional Genomics, University of Adelaide, Urrbrae, SA, Australia*
[5]*CSIRO Plant Industry, Black Mountain Laboratories, Canberra, ACT, Australia*
[6]*CSIRO Food Future Flagship, Black Mountain Laboratories, Canberra, ACT, Australia*
[7]*Department of Primary Industries, Victorian AgriBiosciences Centre, Bundoora, Vic., Australia*
[8]*Bioplatforms Australia, Macquarie University, North Ryde, NSW, Australia*

Received 31 January 2012;
revised 19 April 2012;
accepted 20 April 2012.
**Correspondence* (Tel +612 9850 1174;
fax +612 9850 6200;
email afitzgerald@bioplatforms.com)

Keywords: bread wheat, whole genome sequencing, single nucleotide polymorphisms.

Summary

The large and complex genome of wheat makes genetic and genomic analysis in this important species both expensive and resource intensive. The application of next-generation sequencing technologies is particularly resource intensive, with at least 17 Gbp of sequence data required to obtain minimal (1×) coverage of the genome. A similar volume of data would represent almost 40× coverage of the rice genome. Progress can be made through the establishment of consortia to produce shared genomic resources. Australian wheat genome researchers, working with Bioplatforms Australia, have collaborated in a national initiative to establish a genetic diversity dataset representing Australian wheat germplasm based on whole genome next-generation sequencing data. Here, we describe the establishment and validation of this resource which can provide a model for broader international initiatives for the analysis of large and complex genomes.

Introduction

On a 4-year average (2006–2010), Australia was the ninth largest producer of wheat, 16th largest consumer of wheat, and fifth largest exporter of wheat in the world (United States Department of Agriculture Foreign Agricultural Service, 2010). Intensive breeding has led to significant increases in yield and productivity, although various biotic and abiotic stresses can cause major yield loss in Australia. While many crops such as rice and maize have benefited from advanced genomic tools and complete genome sequences, the size and complexity of the wheat genome have limited genomic applications for wheat improvement. The lack of supportive knowledge on crop genomics is a serious impediment towards tapping potential biotechnological tools for crop improvement. Hence, concerted efforts are required to characterize genetic diversity in Australian bread wheat varieties to enable the generation of superior genotypes underpinning crop improvement, productivity and resilience.

Wheat has a very large genome, estimated to be 17 Gbp in size (Paux *et al.*, 2008). The large size of the wheat genome is in part attributable to being an allohexaploid, meaning that it contains three distinct genomes. The diploid donor species, AA, BB and DD are thought to have diverged between 2.5 and 4.5 MYA and combined to produce *Triticum aestivum* in two distinct hybridization events. First, *Triticum urartu* (AA) and an unknown relative of *Aegilops speltoides* (BB) are believed to have produced the tetraploid *Triticum turgidum* ssp. *dicoccoides* around 0.2–0.5 MYA (Huang *et al.*, 2002). This was followed by hybridization with *Aegilops tauschii* (DD) around 8500 years ago to produce the hexaploid *T. aestivum* (Kihara, 1944; McFadden and Sears, 1946). In addition to polyploidy, the wheat genome has experienced significant proliferation of repetitive elements, resulting in a composition of between 75% and 90% repetitive DNA sequences (Flavell *et al.*, 1977; Wanjugi *et al.*, 2009). This level of complexity hinders the development and application of genomic tools for wheat crop improvement.

The application of molecular markers to advance cereal breeding is now well established (Edwards, 2007; Edwards and Batley, 2008; Gupta *et al.*, 2001). Modern cereal breeding is dependent on molecular markers for the rapid and precise analysis of germplasm, trait mapping and marker assisted selection (Lai *et al.*, 2012b). Molecular markers can be used to select parental genotypes in breeding programmes, eliminate linkage drag in backcrossing and select for traits that are difficult to measure using phenotypic assays (Duran *et al.*, 2010). Molecular markers have many other uses in genetics, such as the discovery of alleles associated with agronomic traits, verification of variety distinctness, uniformity and stability assessment, and inferences of population history (Duran *et al.*, 2009b). Furthermore, molecular markers are invaluable as a tool for genome mapping in all systems, offering the potential for generating very high-density genetic maps that can be used to develop haplotypes for genes or regions of interest (Duran *et al.*, 2009a;

Rafalski, 2002). SNPs represent the most frequent type of genetic polymorphism and may therefore provide a high density of markers and therefore increased mapping resolution near a locus of interest (Duran *et al.*, 2010).

In November 2003, a USDA-NSF funded international workshop of wheat geneticists and sequencing specialists identified the first objectives towards sequencing the hexaploid wheat genome, that is, physical mapping and assessment of sequencing strategies. To capitalize on the momentum of this workshop, the International Wheat Genome Sequencing Consortium (IWGSC, http://www.wheatgenome.org) was established in January 2005 with the goal of coordinating the international effort to build the foundation for and lead the sequencing of the bread wheat genome. The IWGSC has achieved success in engaging countries worldwide in tackling the wheat genome through an approach to sequence flow-sorted chromosomes and thus reduce the complexity of the genome, followed by the construction of DNA BAC libraries from the purified chromosome arm for detailed analysis (Doležel *et al.*, 2007; Molnár *et al.*, 2011; Safar *et al.*, 2004; Šafář *et al.*, 2010). The first BAC library has been used successfully in a project to establish a sequence-ready physical map of chromosome 3B, the largest wheat chromosome (2× the rice genome), and was published by Paux *et al.* (2008).

In parallel to the BAC-based analysis of the wheat genome, the larger data volumes from the Illumina sequencing platform, combined with advanced bioinformatics provide the potential to gain insight into complex plant genomes (Berkman *et al.*, 2012a; Lee *et al.*, 2012; Marshall *et al.*, 2010). This data has been applied for rapid genome sequencing (Batley and Edwards, 2009; Edwards and Batley, 2010; Imelfort and Edwards, 2009) as well as to discover very large numbers of genome-wide SNPs (Imelfort *et al.*, 2009). More than one million SNPs have been identified between six inbred maize lines (Lai *et al.*, 2010). This study also identified a large number of presence/absence variations, which may be associated with heterosis in this species. More recently, Allen *et al.* (2011) identified 14 078 putative SNPs in 6255 distinct reference sequences with Illumina GAIIx data from wheat lines Avalon, Cadenza, Rialto, Savannah and Recital. The validation rate from a subset of 1659 was 67%. A pipeline package called AGSNP has been applied to identify SNPs between two accessions of one of the diploid progenitors of bread wheat, *A. tauschii* (Luo *et al.*, 2009). Roche 454 sequencing of *A. tauschii* accession AL8/78 has since been combined with Applied Biosystems SOLiD sequencing of genomic DNA and cDNA from *A. tauschii* accession AS75 using AGSNP to identify a total of 497 118 candidate *A. tauschii* SNPs (You *et al.*, 2011).

Given the progress in the application of next-generation sequencing in other complex crop species, there is a significant opportunity to apply these approaches to understand genomic diversity in hexaploid bread wheat. However, the size of the genome presents challenges in terms of meeting the cost and sequencing throughput requirements. A large national initiative was established in Australia in 2010, to coordinate diverse wheat genetic and genomic activities and establish a resource for Australian crop improvement. With investment from Bioplatforms Australia and support from the Australian Genome Research Facility, the consortium has succeeded in generating between 5× and 10× coverage of 16 varieties chosen to represent the diversity of Australian wheat germplasm. This resource promises to be a foundation for SNP discovery, supporting Australian wheat crop improvement in the coming decades and provides a model for other national and international crop genomics initiatives. Here, we describe the coordinated development of this resource together with preliminary analysis and quality assessment of the data.

Dataset design and generation

Method for selection

The wheat cultivars were chosen according to three criteria: they represent genetic diversity and have an economic impact in Australia; they are used in building genetic resources such as genetic populations or biotechnologies (parental lines and transformation); and they are key varieties that are both internationally studied and relevant to research in Australia (Rocca-Serra *et al.*, 2010; Sansone *et al.*, 2012; Taylor *et al.*, 2008). After categorization and ranking of 46 suggested lines based on input from breeders, researchers and other stakeholders, a total of 16 lines were selected (Table 1).

Five to ten plants of each line were grown in a growth chamber or glasshouse. DNA was extracted from leaf samples of each plant using a standard phenol/chloroform method as described in Pallotta *et al.* (2000). Each plant was fingerprinted using a set of 10–20 molecular markers to verify the consistency of the germplasm with known genetic resources: parental lines were compared with derived segregating populations; highly variable microsatellites markers were also used to discriminate known versions of some cultivars. Several biotypes have been previously identified within cv. Wyalkatchem. We chose a biotype that has been used as a recurrent parent in backcrossing projects at the University of Adelaide and has been characterized for a number of known loci (seed and information kindly provided by Howard Eagles). For other varieties, such as Chara and Baxter, known to have biotypes, selection was based on an individual plant used in generating the mapping populations currently being utilized.

The sequencing was performed on one DNA sample from a single plant with the corresponding consistent fingerprint. The same plant was seed multiplied by bagging each spike to ensure pure self-crossing. The seed stocks are available through the Australian Pre-breeding Alliance database at the Australian Winter Cereals Collection (http://www2.dpi.qld.gov.au/extra/asp/AusPGRIS/). Each strain will be designated with the cultivar name followed by the BPA suffix (Table 1).

Data generation

Wheat genomic DNA was assessed for quality using the NanoDrop ND-1000 spectrophotometer (ThermoScientific, Willmington, DE) and standard agarose gel electrophoresis. DNA libraries for sequencing were prepared using Illumina TruSeq DNA Library Preparation Kits (Cat. No. FC-390-1021, Illumina Inc., San Diego, CA) and associated recommended protocol. The protocol required that 1 μg of input genomic DNA be sheared using the Covaris S2 (Covaris Inc., Woburn, MA) which resulted in a peak fragment size of 200 bp. Fragmented DNA samples then underwent end-repair to generate blunt ends followed by an A-Tailing reaction to create a uniform 3′ overhang. This overhang was used to ligate the Illumina adapters and index sequences required for the sequencing reaction and subsequent variety identification.

The ligated DNA fragments were purified using the Qiagen MinElute Gel Extraction Kit (Cat. No. 28604, Qiagen,

Table 1 Wheat varieties selected for whole genome shotgun sequencing

Wheat variety		A	B	C	Gbp	Pedigree
1	AC Barrie		×	×	181	NEEPAWA/COLUMBUS//BW-90
2	Alsen		×	×	129	ND-674/ND-2710/ND-688
3	Baxter	×	×		135	INIA-66/GAMUT//COOK/4/JUPATECO/3/LERMA-ROJO-64/SONORA-64-A//(SIB)TIMGALEN
4	Chara	×	×		273	BD-225/CD-87
5	Drysdale		×		160	HARTOG*3/QUARRION
6	Excalibur		×		171	RAC-l77(Sr26)/UNICULM-492//RAC-311-S
7	Gladius	×	×		201	RAC-875/KRICHAUFF//EXCALIBUR/KUKRI/3/RAC-875/KRICHAUFF/4/RAC-875//EXCALIBUR/KUKRI
8	H45	×	×		189	KALYANSONA/BLUEBIRD//ANZA*3/WW-80/3/OLYMPIC*2/CIANO-67
9	Kukri		×		173	CO-1213/RAC-549
10	Pastor		×	×	214	PFAU/SERI-82//BOBWHITE
11	RAC875		×		159	RAC-655/3/Sr21/4*LANCE//4*BAYONET
12	VolcaniDDI (V761-28-J4-B2-NZ8†)		×	×	168	BTL/3/NURSIT-163/G-25/M-708
13	Westonia	×	×		123	SPICA/TIMGALEN//TOSCA/3/CRANBROOK//BOBWHITE*2/JACUP
14	Wyalkatchem	×	×		332	MACHETE/3/(84-W-129-504)GUTHA//JACIP*2/11th-ISEPTON-135
15	Xiaoyan 54‡		×	×	243	ST-2422-464/XIAOYAN-86
16	Yitpi	×	×		222	C-8-MMC-8-HMM/FRAME

Varieties were selected on the basis of A: genetic diversity, B: availability of derived genetic resources and C: potential international interest. Pedigrees: 1–14, 16 as documented by the Genetic Resources Information System for Wheat and Triticale (http://wheatpedigree.net) and 15 selection history as per Grama *et al.* (1984) and Grama *et al.* (1987).

†Selection history of line sequenced.

‡Selection from Xiaoyan 6.

Germantown, MD) and size selected by agarose gel to isolate fragments in the range of 300–400 bp. These fragments were amplified with ten cycles of PCR according to the TruSeq protocol, and the size and concentration of the final library were measured using a Bioanalyser DNA 1000 chip and fluorimetry (PicoGreen QuantIT assay, Molecular Probes Inc., Eugene, OR).

Completed libraries were denatured and diluted to 7 pM for clonal bridge amplification on the Illumina cBot. To obtain the required 10× raw coverage (calculated at 160 Gbp of data) for each variety, on average, three varietal DNA libraries were sequenced across two Illumina HiSeq flowcells according to manufacturer's protocols using multiplex indexes. Base-calling was processed with Illumina RTA software v1.10.36 (currently 1.12, Illumina Inc.). De-multiplexing and conversion to FastQ format were performed with CASAVA v1.7 (Illuminia Inc.) and the later runs with the upgraded v1.8.

Data quality control

To ensure high-quality reads are available for downstream analyses, sequenced datasets were subjected to a series of processing steps. First, poor-quality reads were removed using the following criteria: (i) contain ≥5% of bases as ambiguous calls (Ns), (ii) consist entirely of adenosines (poly A), (iii) the base quality for ≥50% of bases is lower than 7, (iv) both reads of the mate-pair are identical (PCR duplications), (v) after adaptor trimming reads are <50% of initial length. Next, possible mate-pair overlaps were evaluated, where the last ten bases of the first mate were aligned onto the second mate. If a perfect alignment was found, then the alignment was extended allowing up to 10% sequence divergence. Overlapping mate-pairs were joined as extended single end reads, and both fasta and fastq files for these sequences were generated. Finally, poor-quality bases ($Q \leq 10$) at the 3′ end of the reads were trimmed and reads with ≥50% of initial length were removed.

Australian wheat varieties database, data sharing and accessibility

The Australian wheat varieties sequencing data have been curated using an experimental metadata relational database based on Investigation Study Assay (ISA) infrastructure (Rocca-Serra *et al.*, 2010; http://www.isa-tools.org/). The database captures relevant metadata and can be searched and browsed using a web-based interface that also provides links directly to the externally stored data files for download.

We have used ISA-tab format and the minimum information framework defined by (MIBBI) Minimum Information for Biological and Biomedical Investigations (Taylor *et al.*, 2008) to report necessary metadata to facilitate reproducibility and reuse of this Australian wheat varieties reference dataset. Adoption of the ISA infrastructure to curate this data connects this initiative to a growing ISA data commons that promote public data sharing between diverse research domains and maximizes the potential benefit of this dataset to the greater scientific community (Sansone *et al.*, 2012; http://isacommons.org/).

All raw and quality controlled data are publicly available through the Australian wheat varieties database (http://www.bioplatforms.com.au/datasets/wheat/wheat-sequencing/variety-sequencing). We request that analysed data be contributed back to the database to maximize the usefulness of the resource. Users of this data are required to act responsibly and ethically and to

adhere to the Bioplatforms Australia Data Release Policy (http://www.bioplatforms.com.au/datasets/wheat).

Data validation

SNP concordance with the 9k SNP assay

To assess the utility of the Bioplatforms Australia wheat genomic resource for SNP discovery, the accuracy for genotype-calling at 1600 characterized SNP loci was investigated. The 1600 SNPs were chosen as they had known genotypes in a subset of the sequenced BPA wheat varieties, determined from genotyping of the DNA samples supplied for genomic sequencing using an Illumina 9000-feature Infinium SNP assay developed by the International Wheat SNP Working Group (http://wheat.pw.usda.gov/ggpages/9K_assay_available.html), and genotyping-by-sequencing using transcriptome sequence generated for the varieties.

Genomic sequence for varieties AC Barrie, Baxter and Chara was used to assess genotype-calling accuracy at the 1600 validated SNP loci. Following quality trimming, the processed raw sequence reads for each variety (corresponding to about 6×, 7× and 11× genome coverage for AC Barrie, Baxter and Chara, respectively) were mapped against reference sequence for the 1600 SNPs using BWA software (Li and Durbin, 2009). SNP genotypes were called using custom scripts when the minimum coverage at the sequence variant position was seven or ten reads. Assuming a homoeolog-specific reference sequence, a minimum coverage of seven and ten reads at the SNP position corresponds to 87.5% and 97.9% statistical confidence for the genotype call (Galan *et al.*, 2010).

With a minimum coverage of ten reads at the SNP position, genotype calls from the genomic sequence had 91%, 85% and 96% concordance with the validated SNP genotypes for AC Barrie, Baxter and Chara, respectively. However, only 488, 217 and 761 SNP genotypes were called from the sequence data in total, respectively. When the minimum coverage at the SNP position was reduced to seven reads, 637, 392 and 844 genotypes were called with similar concordance, respectively. These results indicate that the Bioplatforms Australia genomic resource can be used to reliably call SNP genotypes with high statistical confidence at homoeolog-specific loci and imply its utility for *de novo* SNP discovery. Examination of the genomic distribution of SNPs (which have been genetically mapped, $n = 901$) across chromosome groups and sub-genomes showed no bias for SNP discovery across the wheat genome (Table 2). The results further suggest that at 6–10× genome coverage, reliable SNP discovery and genotype-calling can be typically achieved for about 24%–52% of SNPs and indicate that increased genome coverage would further improve the accuracy of SNP discovery. This observation is consistent with the requirement of the SNP discovery pipeline for a minimum coverage of seven reads at a sequence variant position for genotype-calling. A detailed description for *de novo* SNP discovery using the Bioplatforms Australia genomic resource will be published elsewhere.

Table 2 Genomic distribution of SNPs used for validation

	Sub-genome		
Chromosome group	A (%)	B (%)	D (%)
1	75 (88)	54 (87)	10 (89)
2	57 (92)	98 (91)	12 (73)
3	52 (95)	60 (89)	3 (100)
4	46 (87)	26 (90)	7 (100)
5	64 (88)	101 (94)	14 (100)
6	56 (89)	73 (95)	9 (85)
7	47 (92)	33 (91)	4 (100)

Per cent in bracket indicates average genotype concordance between known and predicted genotypes for AC Barrie, Baxter and Chara.

SNP discovery by genome mapping

Where reference genomic sequence assemblies are available, it is possible to predict genomic SNPs by mapping paired sequence reads from whole genome shotgun sequencing to the reference. Consistent variety-specific sequence variation within the aligned reads is indicative of variety-specific SNPs. As wheat varieties are predominantly homozygous across their genomes, relatively low coverage is required compared to similar approaches using heterozygous populations. However, SNP discovery can be confounded by the presence of multiple genomes, and reference sequences representing each of the sub-genomes are required to differentiate between homologous (inter genomic) and homoeologous (inter varietal) SNPs. We have developed SGSautoSNP (Second Generation Sequencing autoSNP) software specifically to predict SNPs from whole genome Illumina shotgun sequence data from homozygous polyploid species, and this has been successfully applied to identify more than 1.5 million SNPs across the canola genome with an accuracy of >96% (D. Edwards, personal communication). Reference genomic templates are currently only available for wheat group 7 chromosomes (Berkman *et al.*, 2011, 2012b; Lai *et al.*, 2012a). To assess whether these could be used for genome mapping-based SNP discovery, whole genome shotgun data for an initial four Australian varieties were mapped using SOAP (Li *et al.*, 2008). No pre-filtering of the data was performed with the exception of duplicate read removal. This initial test identified more than 900 000 SNPs between four Australian varieties along this chromosome group and suggests that this approach could be applied to the complete wheat variety dataset. SNP density varied between the genomes, with fewer SNPs identified on 7D (0.40 SNPs per Kbp) compared to 7A (1.69 SNPs per Kbp) and 7B (1.39 SNPs per Kbp). These preliminary results are presented within a GBrowse database and are available publicly at http://www.wheatgenome.info.

Conclusions

Through the establishment of a national initiative, we have produced whole genome shotgun data for 16 wheat varieties representing the diversity within Australian cultivated bread wheat. Preliminary validation suggests that this data is suitable for the identification of genome-wide sequence polymorphisms. This data is publicly accessible and presents a valuable resource for wheat crop improvement in both Australia and internationally.

Acknowledgements

This project was funded by Bioplatforms Australia through the Australian Government's Education Investment Fund (EIF) Super

Science Initiative. iVEC for providing access to computational resources and storage.

References

Allen, A.M., Barker, G.L.A., Berry, S.T., Coghill, J.A., Gwilliam, R., Kirby, S., Robinson, P., Brenchley, R.C., D'Amore, R., McKenzie, N., Waite, D., Hall, A., Bevan, M., Hall, N. and Edwards, K.J. (2011) Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). *Plant Biotechnol. J.* **9**, 1086–1099.

Batley, J. and Edwards, D. (2009) Genome sequence data: management, storage, and visualization. *Biotechniques*, **46**, 333–336.

Berkman, P.J., Skarshewski, A., Lorenc, M.T., Lai, K., Duran, C., Ling, E.Y.S., Stiller, J., Smits, L., Imelfort, M., Manoli, S., McKenzie, M., Kubalakova, M., Simkova, H., Batley, J., Fleury, D., Dolezel, J. and Edwards, D. (2011) Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. *Plant Biotechnol. J.* **9**, 768–775.

Berkman, P.J., Lai, K., Lorenc, M.T. and Edwards, D. (2012a) Next generation sequencing applications for wheat crop improvement. *Am. J. Bot.* **99**, 365–371.

Berkman, P.J., Skarshewski, A., Manoli, S., Lorenc, M.T., Stiller, J., Smits, L., Lai, K., Campbell, E., Kubalakova, M., Simkova, H., Batley, J., Dolezel, J., Hernandez, P. and Edwards, D. (2012b) Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theor. Appl. Genet.* **124**, 423–432.

Doležel, J., Kubaláková, M., Paux, E., Bartoš, J. and Feuillet, C. (2007) Chromosome-based genomics in the cereals. *Chromosome Res.* **15**, 51–66.

Duran, C., Appleby, N., Edwards, D. and Batley, J. (2009a) Molecular genetic markers: discovery, applications, data storage and visualisation. *Curr. Bioinform.* **4**, 16–27.

Duran, C., Edwards, D. and Batley, J. (2009b) Molecular marker discovery and genetic map visualisation. In *Bioinformatics, tools and applications* (Edwards, D., Hanson, D. and Stajich, J., eds), pp. 165–190, New York: Springer.

Duran, C., Eales, D., Marshall, D., Imelfort, M., Stiller, J., Berkman, P.J., Clark, T., McKenzie, M., Appleby, N., Batley, J., Basford, K. and Edwards, D. (2010) Future tools for association mapping in crop plants. *Genome*, **53**, 1017–1023.

Edwards, D. (2007) Bioinformatics and plant genomics for staple crops improvement. In *Breeding Major Food Staples* (Kang, M.S. and Priyadarshan, P.M., eds), pp. 93–106, Ames: Blackwell.

Edwards, D. and Batley, J. (2008) Bioinformatics: fundamentals and applications in plant genetics, mapping and breeding. In *Principles and Practices of Plant Genomics* (Kole, C. and Abbott, A.G., eds), pp. 269–302, Enfield: Science Publishers, Inc.

Edwards, D. and Batley, J. (2010) Plant genome sequencing: applications for crop improvement. *Plant Biotechnol. J.* **7**, 1–8.

Flavell, R.B., Rimpau, J. and Smith, D.B. (1977) Repeated sequence DNA relationships in four cereal genomes. *Chromosoma*, **63**, 205–222.

Galan, M., Guivier, E., Caraux, G., Charbonnel, N. and Cosson, J.-F. (2010) A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics*, **11**, 296.

Grama, A., Gerecther-Amitai, Z.K., Blum, A. and Rubenthaler, G. L. (1984) Breeding bread wheat cultivars for high protein content by transfer of protein genes from Triticum dicoccoides. *In Cereal Grain Protein Improvement*, pp. 145–153. Vienna: International Atomic Energy Agency, Series 681-E.

Grama, A., Porter, N.G. and Wright, D.S.C. (1987) Hexaploid wild emmer wheat derivatives grown under New Zealand conditions 2. Effect of foliar urea sprays on plant and grain nitrogen and baking quality. New Zealand *J. Agri. Res.* **30**, 45–51.

Gupta, P.K., Roy, J.K. and Prasad, M. (2001) Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Curr. Sci.* **80**, 524–535.

Huang, S., Sirikhachornkit, A., Su, X.J., Faris, J., Gill, B., Haselkorn, R. and Gornicki, P. (2002) Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat. *Proc. Natl Acad. Sci. USA*, **99**, 8133–8138.

Imelfort, M. and Edwards, D. (2009) De novo sequencing of plant genomes using second-generation technologies. *Brief. Bioinform.* **10**, 609–618.

Imelfort, M., Duran, C., Batley, J. and Edwards, D. (2009) Discovering genetic polymorphisms in next-generation sequencing data. *Plant Biotechnol. J.* **7**, 312–317.

Kihara, H. (1944) Discovery of the DD-analyzer, one of the ancestors of vulgare wheats. *Agric. Hortic.* **19**, 2.

Lai, J.S., Li, R.Q., Xu, X., Jin, W.W., Xu, M.L., Zhao, H.N., Xiang, Z.K., Song, W.B., Ying, K., Zhang, M., Jiao, Y.P., Ni, P.X., Zhang, J.G., Li, D., Guo, X.S., Ye, K.X., Jian, M., Wang, B., Zheng, H.S., Liang, H.Q., Zhang, X.Q., Wang, S.C., Chen, S.J., Li, J.S., Fu, Y., Springer, N.M., Yang, H.M., Wang, J.A., Dai, J.R., Schnable, P.S. and Wang, J. (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat. Genet.* **42**, 1027–1030.

Lai, K., Berkman, P.J., Lorenc, M.T., Duran, C., Smits, L., Manoli, S., Stiller, J. and Edwards, D. (2012a) WheatGenome.info: An integrated database and portal for wheat genome information. *Plant Cell Physiol.* **53**, 1–7.

Lai, K., Lorenc, M.T. and Edwards, D. (2012b) Genomic databases for crop improvement. *Agronomy*, **2**, 62–73.

Lee, H., Lai, K., Lorenc, M.T., Imelfort, M., Duran, C. and Edwards, D. (2012) Bioinformatics tools and databases for analysis of next generation sequence data. *Brief. Funct. Genomics*, **2**, 12–24.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, R.Q., Li, Y.R., Kristiansen, K. and Wang, J. (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.

Luo, M.C., Deal, K.R., Akhunov, E.D., Akhunova, A.R., Anderson, O.D., Anderson, J.A., Blake, N., Clegg, M.T., Coleman-Derr, D., Conley, E.J., Crossman, C.C., Dubcovsky, J., Gill, B.S., Gu, Y.Q., Hadam, J., Heo, H.Y., Huo, N., Lazo, G., Ma, Y., Matthews, D.E., McGuire, P.E., Morrell, P.L., Qualset, C.O., Renfro, J., Tabanao, D., Talbert, L.E., Tian, C., Toleno, D.M., Warburton, M.L., You, F.M., Zhang, W. and Dvorak, J. (2009) Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. *Proc. Natl Acad. Sci. USA*, **106**, 15780–15785.

Marshall, D., Hayward, A., Eales, D., Imelfort, M., Stiller, J., Berkman, P., Clark, T., McKenzie, M., Lai, K., Duran, C., Batley, J. and Edwards, D. (2010) Targeted identification of genomic regions using TAGdb. *Plant Methods*, **6**, 19.

McFadden, E. and Sears, E. (1946) The origin of *Triticum spelta* and its free-threshing hexaploid relatives. *J. Hered.* **37**, 81–89.

Molnár, I., Kubaláková, M., Šimková, H., Cseh, A., Molnár-Láng, M. and Doležel, J. (2011) Chromosome Isolation by Flow Sorting in Aegilops umbellulata and Ae. comosa and Their Allotetraploid Hybrids Ae. biuncialis and Ae. geniculata. *PLoS ONE*, **6**, e27708.

Pallotta, M.A., Graham, R.D., Langridge, P., Sparrow, D.H.B. and Barker, S.J. (2000) RFLP mapping of manganese efficiency in barley. *Theor. Appl. Genet.* **101**, 1100–1108.

Paux, E., Sourdille, P., Salse, J., Saintenac, C., Choulet, F., Leroy, P., Korol, A., Michalak, M., Kianian, S., Spielmeyer, W., Lagudah, E., Somers, D., Kilian, A., Alaux, M., Vautrin, S., Berges, H., Eversole, K., Appels, R., Safar, J., Simkova, H., Dolezel, J., Bernard, M. and Feuillet, C. (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. *Science*, **322**, 101–104.

Rafalski, A. (2002) Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* **5**, 94–100.

Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., Field, D., Harris, S., Hide, W., Hofmann, O., Neumann, S., Sterk, P., Tong, W. and Sansone, S.-A. (2010) ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, **26**, 2354–2356.

Safar, J., Bartos, J., Janda, J., Bellec, A., Kubalakova, M., Valarik, M., Pateyron, S., Weiserova, J., Tuskova, R., Cihalikova, J., Vrana, J., Simkova, H., Faivre-Rampant, P., Sourdille, P., Caboche, M., Bernard, M., Dolezel, J. and Chalhoub, B. (2004) Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. *Plant J.* **39**, 960–968.

Šafář, J., Šimková, H., Kubaláková, M., Číhalíková, J., Suchánková, P., Bartoš, J. and Doležel, J. (2010) Development of chromosome-specific BAC resources for genomics of bread wheat. *Cytogenet. Genome Res.* **129**, 211–223.

Sansone, S.-A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., Begley, K., Booth, T., Bougueleret, L., Burns, G., Chapman, B., Clark, T., Coleman, L.-A., Copeland, J., Das, S., de Daruvar, A., de Matos, P., Dix, I., Edmunds, S., Evelo, C.T., Forster, M.J., Gaudet, P., Gilbert, J., Goble, C., Griffin, J.L., Jacob, D., Kleinjans, J., Harland, L., Haug, K., Hermjakob, H., Sui, S.J.H., Laederach, A., Liang, S., Marshall, S., McGrath, A., Merrill, E., Reilly, D., Roux, M., Shamu, C.E., Shang, C.A., Steinbeck, C., Trefethen, A., Williams-Jones, B., Wolstencroft, K., Xenarios, I. and Hide, W. (2012) Toward interoperable bioscience data. *Nat. Genet.* **44**, 121–126.

Taylor, C.F., Field, D., Sansone, S.-A., Aerts, J., Apweiler, R., Ashburner, M., Ball, C.A., Binz, P.-A., Bogue, M., Booth, T., Brazma, A., Brinkman, R.R., Michael Clark, A., Deutsch, E.W., Fiehn, O., Fostel, J., Ghazal, P., Gibson, F., Gray, T., Grimes, G., Hancock, J.M., Hardy, N.W., Hermjakob, H., Julian, R.K., Kane, M., Kettner, C., Kinsinger, C., Kolker, E., Kuiper, M., Novere, N.L., Leebens-Mack, J., Lewis, S.E., Lord, P., Mallon, A.-M., Marthandan, N., Masuya, H., McNally, R., Mehrle, A., Morrison, N., Orchard, S., Quackenbush, J., Reecy, J.M., Robertson, D.G., Rocca-Serra, P., Rodriguez, H., Rosenfelder, H., Santoyo-Lopez, J., Scheuermann, R.H., Schober, D., Smith, B., Snape, J., Stoeckert, C.J., Tipton, K., Sterk, P., Untergasser, A., Vandesompele, J. and Wiemann, S. (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* **26**, 889–896.

Wanjugi, H., Coleman-Derr, D., Huo, N., Kianian, S., Luo, M., Wu, J., Anderson, O. and Gu, Y. (2009) Rapid development of PCR-based genome-specific repetitive DNA junction markers in wheat. *Genome*, **52**, 576–587.

You, F., Huo, N., Deal, K., Gu, Y., Luo, M.-C., McGuire, P., Dvorak, J. and Anderson, O. (2011) Annotation-based genome-wide SNP discovery in the large and complex Aegilops tauschii genome using next-generation sequencing without a reference genome sequence. *BMC Genomics*, **12**, 59.